DOCUMENT RESUME

ED 262 063 TM 850 538

AUTHOR Jolly, S. Jean; And Others
TITLE The Effect of Test Speededness and Random Guessing on the Validity of Reading Comprehension Scores.
PUB DATE Apr 85
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Elementary Secondary Education; *Guessing (Tests); Item Analysis; *Reading Comprehension; Reading Rate; *Reading Tests; Testing Problems; *Test Length; Test Results; *Test Validity; *Timed Tests
IDENTIFIERS Palm Beach County Schools FL; *Stanford Achievement Tests

ABSTRACT

Scores from the Stanford Achievement Tests administered to 50,000 students in Palm Beach County, Florida, were studied in order to determine whether the speeded nature of the reading comprehension subtest was related to inconsistencies in the score profiles. Specifically, the probable effect of random guessing was examined. Reading scores were considerably higher or lower than corresponding scores on other subtests, especially in grades 4-7. Item analyses were performed on classroom, school, and district-level test results. Examination of completion rates suggested that deviant reading comprehension performance (as compared to other subtests) was related to unusually high or low completion rates. Further analyses indicated that the students who finished the most reading items were not the highest achievers on the other tests. Apparently the speededness of the reading test encouraged random guessing. Norms were made available for the first 40 items of the reading test, thus reducing it from 60 items to 40. Scores on the 40-item test correlated more positively with the other subtests. Moreover, correlations between completion rate and performance on the other subtests were weak. The shorter test resolved the problems associated with the longer test. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262063

The Effect of Test Speededness and Random Guessing on the Validity of Reading Comprehension Scores

S. Jean Jolly, Robert Johnson, Bernard Jones, Jose Abalos
Palm Beach County School District

A paper presented at the Annual Meeting of the
American Educational Research Association
Chicago, April 1985

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Jolly, S. Jean

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 538

INTRODUCTION

Davis (1972) maintains that there are two dimensions of reading ability: power, the ability to comprehend and apply rather difficult textual material in generous time limits; and speed, the rate of comprehension of relatively easy material. Depending on the time limits established by the test publisher, standardized tests of reading comprehension may or may not involve speed as well as power. In the case of a highly speeded test, the reading comprehension score will reflect reading rate as well as power of comprehension.

After reviewing test speededness studies dating back to 1941, Rindler (1979) concluded that there is neither a strong or consistent relationship between speed and power. Using carefully constructed instrumentation, Bloomers and Lindquist (1944) found the correlation between speed and power of comprehension to be only .30. Moreover, Davidson and Carroll (1945) presented evidence that speed scores were linearly independent of power scores. They also found that scores obtained from timed tests were factorially complex, having loadings on both speed and power dimensions of ability.

The standardized norm-referenced test batteries typically administered in school systems, however, do not have separate measures for speed and power. They provide only one measure of reading comprehension, which may or may not involve speed, depending on the restrictiveness of the time limits. Myers (1960) surmised that time limits are used to maximize the usefulness of a class period in order to ensure financial success of tests. Similarly, Morrison (1960) concluded that time limits were used for practical, rather than empirical, reasons. Helmstadtler and Ortmeyer (1953)

recommended that evaluation of a test should include precise knowledge of the relative contributions of speed and power to test scores, and Stafford (1971) called for publishers to report speededness quotients. Indeed, Morrison (1960, p. 232) argued that "time limits cannot be treated casually because they may produce a change in the determination of test scores so that something different from that intended by the test constructor is being measured."

It is generally agreed that restrictive time limits may cause random marking of answers, with guessing increasing as time runs out (Kendall, 1964). This random marking increases the errors of measurement present in the test scores (Lord, 1964). In a study of the effect of time limits on the behavior of test takers, Mollenkopf (1960) found that subjects who realized that they would be unable to finish a test often marked answers randomly, with chance making a significant difference in scores.

Swineford (1956) considers a test to be unspeeded if virtually all subjects attempted 75 percent of the items and at least 80 percent responded to the last item. The Stanford Achievement Test, Seventh Edition (SAT/7), Intermediate I Form E Reading Comprehension subtest (a 60-item measure with a 30-minute time limit) does not meet these criteria. Only 71 percent of the national sample at grade 4 reached item 45, and only 54 percent responded to the last item (J. M. Lenke, personal communication, September 1983).

The SAT/7 assesses reading comprehension through two approaches: as it relates to the type of material read and as it relates to the type of question asked (Gardner, Madden, Rudman, Karlsen, Merwin, Collins, & Collins, 1984). No mention is

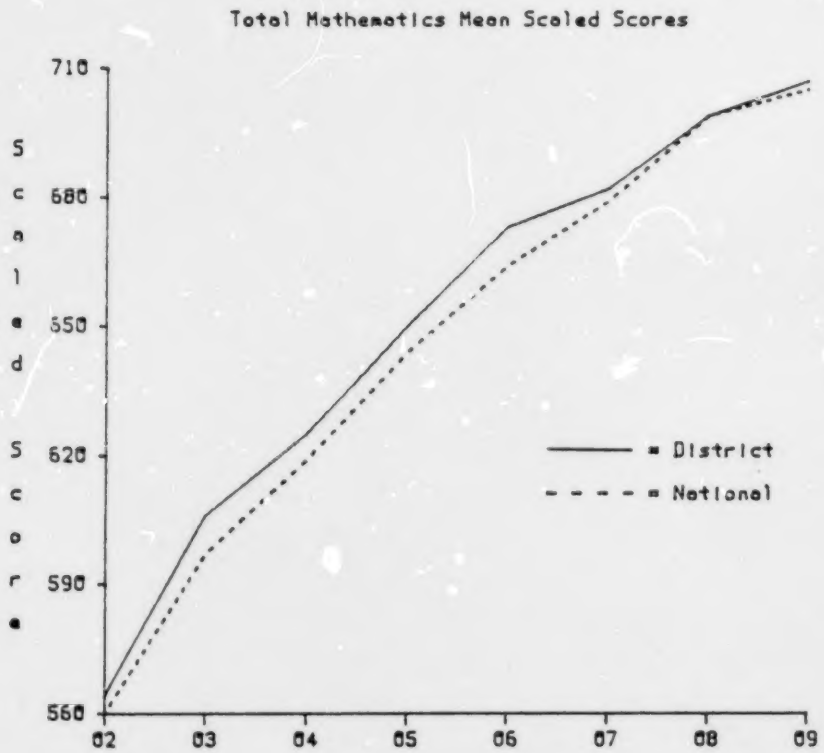
made of the measurement of rate in any of the published materials (Karlsen, 1982a; Karlsen, 1982b; Gardner, Rudman, Karlsen, & Merwin, 1981a and 1981b).

The SAT/7 was administered to approximately 50,000 Palm Beach County, Florida, students in grades 2-9 during April, 1983. As with previous standardized tests, performance on the Stanford Achievement Test for students in Palm Beach County was generally near the national average for all subtests (see Figure 1 for examples). An obvious exception, however, was noted in the Reading Comprehension subtest: an irregularity existed in the growth curve in scaled scores for county-wide performance (see Figure 2). Here the dip from grade 3 (Primary 3 Form E) to grade 4 (Intermediate I Form E) was obvious, with depressed performance observed in grades 4 through 7, returning to near the national average for grades 8 and 9. This atypical growth pattern prompted an investigation into Reading Comprehension scores, focused on grade 4 (Intermediate I Form E) where the dip was first apparent.

This study was an effort to determine whether the speeded nature of the SAT/7 Reading Comprehension subtest (Intermediate I, Form E) was related to the inconsistencies noted in Palm Beach County's student, classroom, school and district SAT/7 score profiles. Specifically, the study focused on the probable effect of random guessing on the Reading Comprehension score.

Figure 1

Palm Beach County Mean Scaled Scores Across Grades
Stanford Achievement Test, 1983



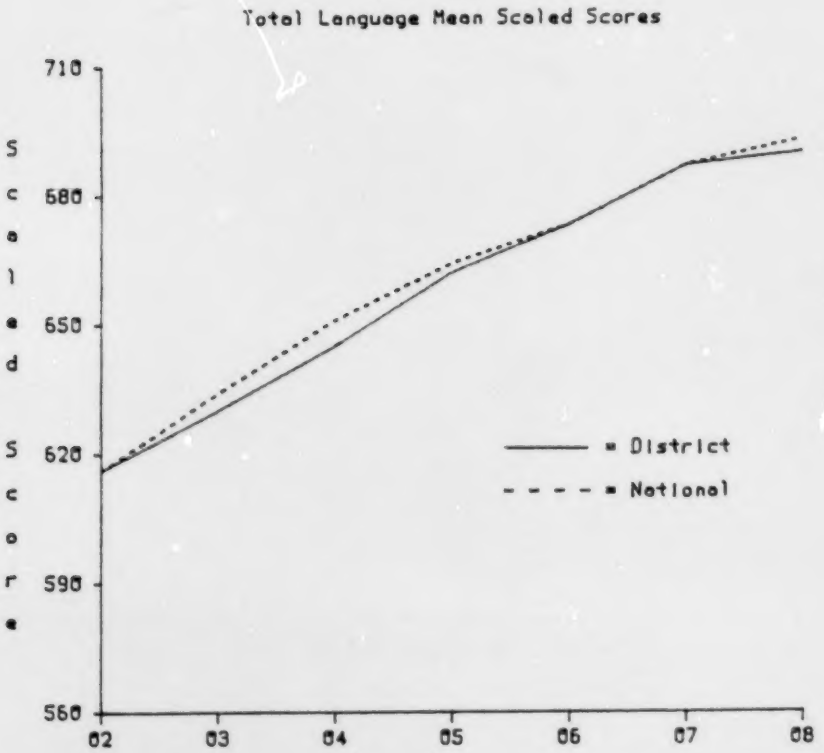
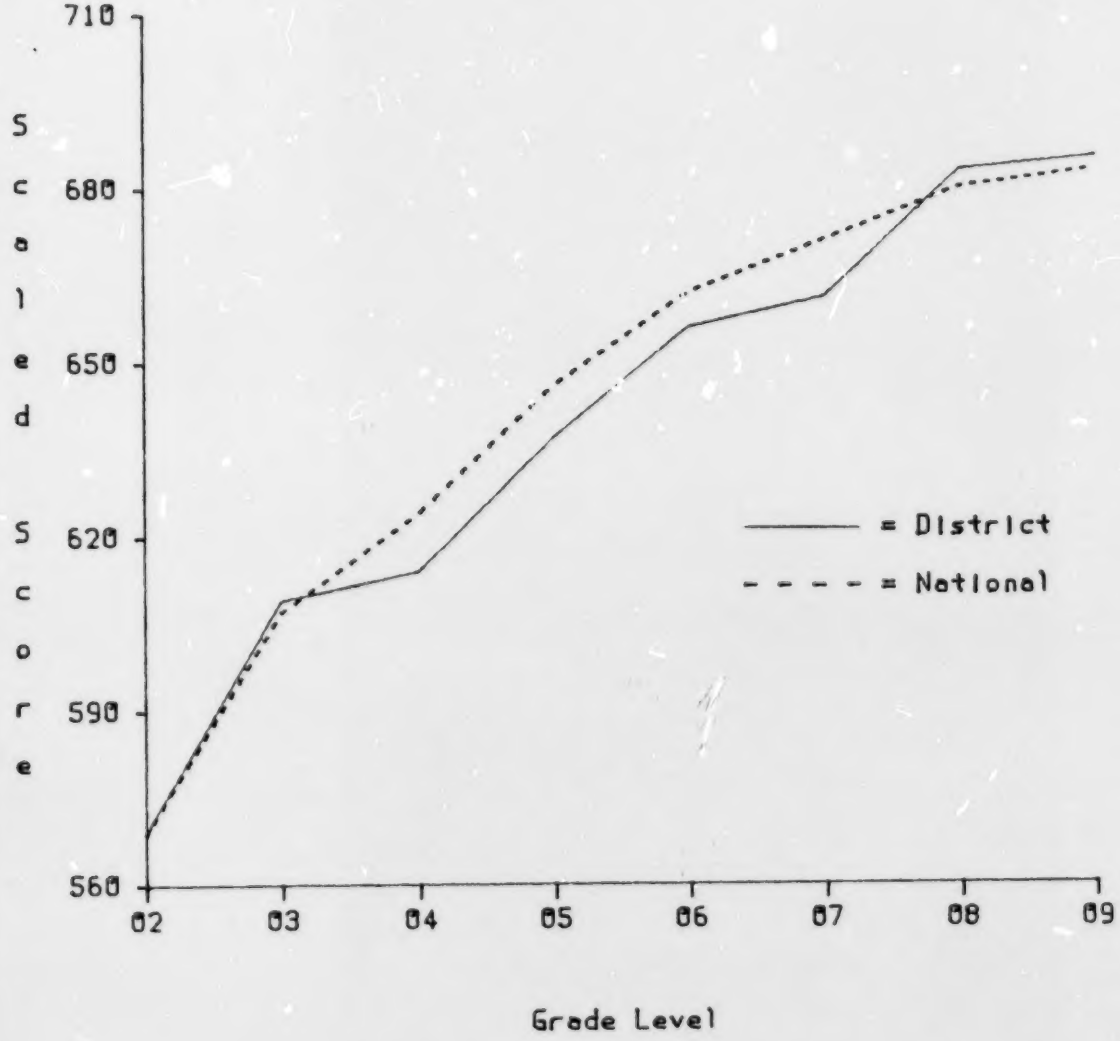


Figure 2

Reading Comprehension Mean Scaled Scores



PROCEDURES AND RESULTS

Analysis of Classroom Performance

The initial investigation of irregular performance at grade 4 consisted of the examination of school and classroom score profiles. The Reading Comprehension scores of some schools and many classrooms were either considerably higher or considerably lower than their corresponding scores on the other SAT/7 subtests (see Figures 3 and 4 for classroom examples). The Reading Comprehension scores of these schools and classes therefore warranted further investigation, being significantly different than group performance on other subtests.

To further investigate these results, item analyses were performed on the classroom, school, and district levels. The district item analysis (Table 1) showed that 61 percent of the students in the county did not answer the last item (i.e., only 39 percent of the students completed the subtest). The item omit percentages increased gradually from 0 to 15 percent on the first thirty items, increasing rapidly from 16 to 45 percent on items thirty-one to forty-five, and then to 61 percent by item sixty. In contrast, on no other SAT/7 subtests did more than 20 percent of the students omit the final item.

The classroom item analyses evidenced several unexpected results. A number of classrooms characterized by low scores on other SAT/7 subtests had unusually high Reading Comprehension completion rates (i.e., percentage of students marking item 60) on the subtest (see Figure 5 for example). On the other hand, some of those class-

Figure 3

Class Profile of Percentile Rank Across Subtests
(Based on Median Scaled Score)

Class #1

Subtest	Percentile Rank
Reading Comprehension	----------32
Word Study Skills	----------58
Vocabulary	----------68
Listening Comprehension	----------79
Spelling	----------56
Language	----------70
Concepts of Numbers	----------77
Mathematics Computation	----------73
Mathematics Application	----------73
Social Science	----------58
Science	----------66

Figure 4

Class Profile of Percentile Rank Across Subtests
(Based on Median Scaled Score)

Class #2

Subtest	Percentile Rank
Reading Comprehension	----------74
Word Study Skills	----------23
Vocabulary	----------10
Listening Comprehension	----------18
Spelling	----------27
Language	----------24
Concepts of Numbers	----------47
Mathematics Computation	----------39
Mathematics Application	----------20
Social Science	----------47
Science	----------44

Table 1

District Item Analysis
Reading Comprehension, Intermediate I Form E, Grade 4

Item No.	P-Value	Omit Percent	Item No.	P-Value	Omit Percent
1	82	0	31	50	16
2	78	0	32	33	17
3	78	1	33	62	22
4	65	0	34	64	23
5	68	0	35	62	24
6	58	0	36	61	25
7	82	0	37	63	27
8	76	0	38	57	29
9	78	0	39	56	30
10	76	0	40	26	30
11	75	0	41	44	39
12	48	1	42	42	40
13	38	1	43	39	43
14	89	1	44	40	43
15	91	1	45	32	45
16	87	1	46	29	46
17	69	1	47	34	47
18	69	1	48	26	48
19	91	1	49	24	48
20	58	2	50	38	53
21	88	3	51	34	54
22	71	3	52	35	56
23	61	4	53	29	56
24	65	4	54	29	57
25	34	5	55	25	58
26	65	9	56	27	60
27	55	10	57	29	60
28	64	12	58	48	61
29	55	14	59	23	61
30	44	15	60	19	61

Figure 5

Low Scoring Class with High Reading Comprehension Completion*

Class #3

Subtest	Percentile Rank
Reading Comprehension	----------42
Word Study Skills	----------15
Vocabulary	---------9
Listening Comprehension	----------17
Spelling	----------23
Language	----------30
Concepts or Numbers	----------42
Mathematics Computation	----------38
Mathematics Application	----------37
Social Science	----------22
Science	----------23

*Percentage of students marking item 60: 81%
mean last item marked: 59
mean no. items marked: 59

rooms characterized by high scores on the other subtests had unusually low completion rates (see Figure 6 for example).

The analysis of completion rates compared with classroom performance revealed an interesting pattern. Generally low scoring classrooms with unusually high Reading Comprehension scores evidenced high completion rates (see Figure 7 for example). On the other hand, generally high scoring classrooms with low Reading Comprehension scores had low completion rates (see Figure 8 for example). In short, the classroom analysis revealed an apparent relationship between deviant performance on Reading Comprehension, relative to other subtests, and unusually low or high completion rates.

Group Comparisons Based on Completion Rates

The observation that there was an apparent relationship between deviant performance and completion rate led to the supposition that an extraneous factor was affecting the Reading Comprehension score. One method suggested for examining this possibility was to compare the performance on other SAT/7 subtests of students completing Reading Comprehension and those who did not complete the subtest. In this analysis the students were sorted into four groups: those who completed the 60-item subtest (Group 1), and those whose last item marked was 45-59 (Group 2), 34-44 (Group 3), and 1-23 (Group 4). The mean scaled score of each of these groups of students on each SAT/7 subtest was then computed. In Table 2 it can be seen that Group 1, students completing the Reading Comprehension subtest, had a lower mean

Figure 6

High Scoring Class with Low Reading Comprehension Completion*

Class #4

Subtest	Percentile Rank
Reading Comprehension	----------46
Word Study Skills	----------73
Vocabulary	----------70
Listening Comprehension	----------64
Spelling	----------76
Language	----------80
Concepts of Numbers	----------82
Mathematics Computation	----------75
Mathematics Application	----------82
Social Science	----------58
Science	----------67

*Percentage of students marking item 60: 10%
mean last item marked: 43
mean no. items marked: 42

Figure 7

Low Scoring Class with High Reading Comprehension Score and High Completion Rate*

Class #2

Subtest	Percentile Rank
Reading Comprehension	----------74
Word Study Skills	----------23
Vocabulary	----------10
Listening Comprehension	----------18
Spelling	----------27
Language	----------24
Concepts of Numbers	----------47
Mathematics Computation	----------39
Mathematics Application	----------20
Social Science	----------47
Science	----------44

*

Percentage of students marking item 60:	100%
mean last item marked:	60
mean no. items marked:	60

Figure 8

High Scoring Class with Low Reading Comprehension Score and Low Completion Rate*

Class #1

Subtest	Percentile Rank
Reading Comprehension	32
Word Study Skills	58
Vocabulary	68
Listening Comprehension	79
Spelling	56
Language	70
Concepts of Numbers	77
Mathematics Computation	73
Mathematics Application	73
Social Science	58
Science	66

*

Percentage of students marking item 60:	16%
mean last item marked:	54
mean no. items marked:	49

score on other subtests than students in Group 2, who completed fewer items (45-59). For example, students in the 45-59 group scored 4.3 scaled score points higher on Language than students completing the Reading Comprehension subtest. Overall, students in Group 1 scored from 1.9 to 6.7 scaled score points lower on the other subtests than students in Group 2.

This analysis was also conducted on the classroom level by comparing classes with high completion rates to those with lower completion rates. Classes were divided into four approximately equal sized groups: those in which the percentage of students marking the last item was 80-100 (Group 1), 65-79 (Group 2), 51-64 (Group 3), and 0-50 (Group 4). The mean scaled score of each of these groups of classes on each SAT/7 subtest was then computed. In this analysis the same pattern prevailed as found in the student analysis (see Table 3). Classes with the highest percentages of students completing the Reading Comprehension subtest scored lower on the other subtests than did classes with lower completion rates. For example, classes with an 80-100 percent completion rate (Group 1) scored 3.3 scaled score points lower on Language than did classes with a 65-79 percent completion rate (Group 2). Overall, Group 1 classes scored 3.2 to 8.8 scaled score points lower on the other subtests than did Group 2 classes. As in the student analysis, the classes which completed the greatest portion of the Reading Comprehension subtest were not the highest achieving students as measured by other subtests.

Table 2

Student Mean Scaled Score by Completion Group

	Last Reading Comprehension Item Marked						
	Group 1		Group 2		Group 3		Group 4
Subtest	60	Diff.	45-59	Diff.	34-44	Diff.	1-33
Reading Comprehension	640.4	22.2	618.2	21.1	597.1	29.6	567.5
Word Study Skills	617.9	-5.0	622.9	3.1	619.8	11.5	608.3
Vocabulary	630.0	-2.8	632.8	6.8	626.0	13.9	612.1
Listening Comp.	637.0	-2.0	639.0	6.1	632.9	11.2	621.7
Total Listening	632.7	-2.5	635.2	6.4	628.8	12.0	616.8
Spelling	634.3	-2.2	636.5	6.7	629.8	22.5	607.3
Language	632.2	-4.3	636.5	3.8	632.7	14.3	618.4
Total Language	631.9	-3.7	635.6	5.2	630.4	16.6	613.8
Number Concepts	630.3	-4.9	635.2	6.2	629.0	21.1	607.9
Math Computation	626.6	-1.9	628.5	4.4	624.1	13.6	610.5
Math Applications	624.7	-6.7	631.4	4.0	627.4	20.7	606.7
Total Mathematics	626.0	-4.0	630.0	4.8	625.2	16.9	608.3
Social Science	619.7	-2.0	621.7	7.7	614.0	21.0	593.0
Science	623.9	-2.5	626.4	5.4	621.0	18.0	603.0
	N-2015		N-1083		N-1054		N-1033

Table 3

Classroom Mean Scaled Score by Completion Group

	------- Percent of Students Marking Last Read Comp Item -------						
	Group 1		Group 2		Group 3		Group 4
Subtest	80-100	Diff.	65-79	Diff.	51-64	Diff.	0-50
Reading Comprehension	625.1	7.2	617.9	7.8	610.1	14.8	595.3
Word Study Skills	614.8	-7.1	621.9	2.9	619.0	4.2	614.8
Vocabulary	624.4	-7.4	631.8	4.2	627.6	5.5	622.1
Listening Comp.	633.2	-3.8	637.0	4.0	633.0	2.2	630.8
Total Listening	628.2	-5.4	633.6	3.8	629.8	3.9	625.9
Spelling	625.7	-6.2	631.9	.5	631.4	6.9	624.5
Language	628.5	-3.3	631.8	-.9	632.7	4.2	628.5
Total Language	626.6	-4.2	630.8	-.5	631.3	5.3	626.0
Number Concepts	624.9	-7.9	632.8	4.3	628.5	7.4	621.1
Math Computation	622.5	-3.4	625.9	4.8	621.1	-1.2	622.3
Math Applications	619.0	-8.8	627.8	.9	626.9	8.0	618.9
Total Mathematics	621.1	-5.9	627.0	2.8	624.2	4.2	620.0
Social Science	612.9	-4.9	617.8	2.6	615.2	6.7	608.5
Science	619.2	-3.2	622.4	.6	621.8	6.9	614.9
	N-57		N-46		N-49		N-50

Rescoring the Reading Comprehension Subtest

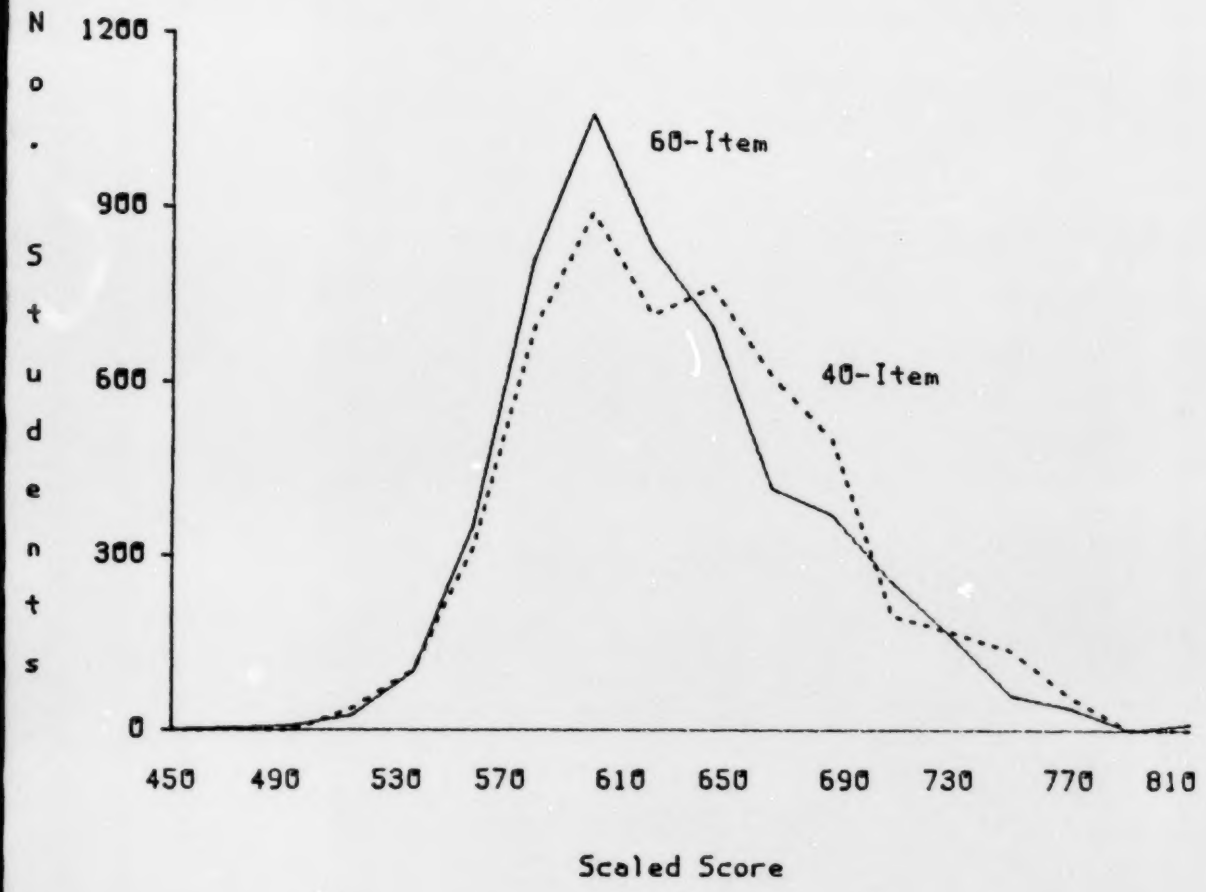
The results of the student and classroom analysis by completion group seemed to indicate that the highly speeded nature of the Reading Comprehension subtest may have encouraged random guessing on the final portion of the test. For this reason the district became interested in the possibility of re-scoring the Reading Comprehension subtest in such a way as to minimize the effect of random guessing. Since the SAT/7 was developed using the Rasch model, the re-scoring of the subtest based on a reduced number of items was possible. The Psychological Corporation, publisher of the SAT/7, provided the district with norms for a reduced set of 40 items. J. J. Fremer (personal correspondence, July 20, 1984) describes the process thusly:

> Since one outcome of the Rasch Model is an ability scale for a test that comes from the difficulties of the items in that test, it is possibile to produce an ability scale for a "short form" of the test, using the difficulties of the items in that short form . . . The norms for the 40-item Reading Comprehension subtest for grades 4 through 9 were developed in a way that parallels the development of the norms for the full-length test. That is, in grade 4, for example, the scaled score distributions for both forms of the Primary 3 and Intermediate 1 levels were combined for common subtests and totals, and norms were developed based on these combined distributions.

The Reading Comprehension subtest was then re-scored on the reduced set of 40 items with new norms, and the results were examined. In response to concerns of shortened test length, the distribution of the 40-item score was inspected for ceiling effects. As seen in Figure 9, the 40-item score distribution was similiar to the distribution of the 60-item subtest. Although the internal consistency estimate (Cronbach's alpha) for the 40-item test (.91) was slightly lower than that for the

Figure 9

Reading Comprehension Scaled Score Distribution (60/40 Item)



60-item test (.95), the latter coefficient was considered spuriously high because of the highly speeded nature of the subtest (Stanley, 1971). The 40-item measure was less speeded than the 60-item subtest, with the district omit percentage being 30 percent at item 40 (compared to 61 percent at item 60).

In order to investigate the relationship of the two Reading Comprehension measures to the other SAT/7 subtests, correlations of the 40- and 60-item scores with other subtest scores were performed for students and for classroom groups. Student 40-item scores correlated more positively with the other subtests than did the corresponding 60-item scores (Table 4). Moreover, the correlations between student completion rates and performance on the other subtests were weak, ranging from .11 to .30 (Table 5).

The classroom level correlations evidenced a similar pattern. Table 6 shows that the 40-item score correlated more highly with other subtests than the 60-item score, with differences that were larger than those observed in the student correlations. The correlations with the 60-item test center in the low .70's, whereas those with 40-items center around .85. Further, the correlation of the class mean last item marked, mean items attempted, and percentage marking last item with other subtest scores was also low, ranging from .01 to .26 (Table 7).

Table 4

Student Correlations of 40 and 60 Item Reading Comprehension Scores With Other Subtests

Subtest	60-Item Reading	40-Item Reading
Reading Comp. (60)	--	.89
Reading Comp. (40)	.89	--
Word Study Skills	.58	.66
Vocabulary	.62	.68
Listening Comp.	.61	.70
Total Listening	.65	.72
Spelling	.64	.68
Language	.64	.71
Total Language	.69	.75
Number Concepts	.59	.65
Math Computation	.49	.53
Math Applications	.61	.69
Total Mathematics	.63	.69
Social Science	.73	.77
Science	.70	.75

N = 5100 (students)

Table 5

Student Correlations of Reading Comprehension Completion With Other Subtests

Subtest	Last Item Marked	No. Items Marked
Reading Comp. (60)	.60	.65
Reading Comp. (40)	.35	.39
Word Study Skills	.11	.13
Vocabulary	.17	.19
Listening Comp.	.17	.20
Total Listening	.18	.21
Spelling	.22	.25
Language	.16	.18
Total Language	.20	.22
Number Concepts	.18	.20
Math Computation	.15	.16
Math Applications	.15	.17
Total Mathematics	.17	.19
Social Science	.27	.30
Science	.22	.26

N = 5100 (students)

Table 6

Classroom Correlations of 40 and 60 Item Reading Comprehension Scores With Other Subtests

Subtest	60-Item Reading	40-Item Reading
Reading Comp. (60)	--	.92
Reading Comp. (40)	.92	--
Word Study Skills	.66	.80
Vocabulary	.68	.81
Listening Comp.	.71	.82
Total Listening	.72	.84
Spelling	.72	.85
Language	.71	.82
Total Language	.74	.86
Number Concepts	.71	.83
Math Computation	.60	.69
Math Applications	.72	.86
Total Mathematics	.71	.84
Social Science	.80	.90
Science	.79	.88

N = 200 (classes)

Table 7

Classroom Correlations of Reading Comprehension Completion With Other Subtests

Subtest	Mean Last Item Marked	Mean No. Items Marked	Percent Answering Last Item
Reading Comp. (60)	.62	.67	.44
Reading Comp. (40)	.34	.38	.12
Word Study Skills	.10	.11	.05
Vocabulary	.12	.15	.01
Listening Comp.	.17	.19	.03
Total Listening	.15	.17	.02
Spelling	.12	.14	-.06
Language	.12	.15	-.05
Total Language	.13	.15	-.05
Number Concepts	.17	.18	.03
Math Computation	.12	.12	-.04
Math Applications	.11	.14	-.05
Total Mathematics	.14	.15	-.02
Social Science	.23	.26	.03
Science	.21	.24	.01

N = 200 (classes)

Application of the 40-Item Norms

The application of the 40-item scoring to the Reading Comprehension subtest produced score changes on the student, classroom, school, and district level. On the student level, the use of the 40-item score created score differences of varying magnitude. Table 8 shows the distribution of change in student scores from the 60-item to the 40-item score. It can be seen that the major score change was a gain of between 6 and 15 scaled score points, with a mean gain of 5.7 scaled score points.

The effect of the 40-item scoring was also examined on the classroom level. The 40-item score was found to be more consistent with classroom scores on the other subtests, particularly in the case of the deviant classrooms found in the initial analysis. For example, Figure 10 shows the same class reported in Figure 3 with a low 60-item Reading Comprehension score (32nd percentile). Use of the 40-item norms resulted in a percentile rank of 51, which was more in line with the rest of the class scores. Similarly, Figure 11 shows the same class reflected in Figure 4 with a high 60-item Reading Comprehension score (74th percentile). The 40-item norms produced a percentile rank of 53, which was more consistent with the other subtest scores.

In order to ascertain the effect of 40-item scoring at the school and district levels, it was necessary to examine results at grades 5-7, where district performance was also depressed. Repeating the investigation outlined above revealed inconsistencies similiar to those observed in grade 4. Use of the publisher developed 40-item norms at these grade levels eliminated the discrepant performance in the same way as it had at grade 4.

Table 8

Student Scaled Score Gain:

Reading Comprehension 40-Item Minus 60-Item

Scaled Score Gain	No. Students	Percent
106 to 119	9	0
96 to 105	10	0
86 to 95	9	0
76 to 85	12	0
66 to 75	25	1
56 to 65	55	1
46 to 55	114	2
36 to 45	160	3
26 to 35	312	6
16 to 25	720	14
6 to 15	1352	27
- 4 to 5	480	10
-14 to - 5	709	14
-24 to -15	593	12
-34 to -25	371	7
-44 to -35	51	1
-54 to -45	2	0

Mean Scaled Score Gain: 5.7

Total N: 4984

Figure 10

Class Profile Showing 40 and 60 Item Reading Comprehension*

Class #1

Subtest	Percentile Rank
Reading Comprehension (60)	----------32
Reading Comprehension (40)	----------51
Word Study Skills	----------58
Vocabulary	----------68
Listening Comprehension	----------79
Spelling	----------56
Language	----------70
Concepts of Numbers	----------77
Mathematics Computation	----------73
Mathematics Application	----------73
Social Science	----------58
Science	----------66

*

Percentage of students marking item 60:	16%
mean last item marked:	54
mean no. items marked:	49

Figure 11

Class Profile Showing 40 and 60 Item Reading Comprehension*

Class #2

Subtest	Percentile Rank
Reading Comprehension (60)	----------74
Reading Comprehension (40)	----------53
Word Study Skills	----------23
Vocabulary	----------10
Listening Comprehension	----------18
Spelling	----------27
Language	----------24
Concepts of Numbers	----------47
Mathematics Computation	----------39
Mathematics Application	----------20
Social Science	----------47
Science	----------44

*Percentage of students marking item 60: 100%
mean last item marked: 60
mean no. items marked: 60

Because of initially deviant 60-item performance, Reading Comprehension school profiles showed certain inconsistencies across grades in some schools. For example, Figure 12 shows a school with unusually low 60-item performance at grade 4 compared to other grade levels (top chart), which was resolved by the use of the 40-item norms (bottom chart).

Finally, the district mean scaled scores obtained from the application of the 40-item norms at grades 4-7 produced a growth curve more consistent with expectation (see Figure 13). District performance exhibited smooth growth across grade levels and remained near the national average.

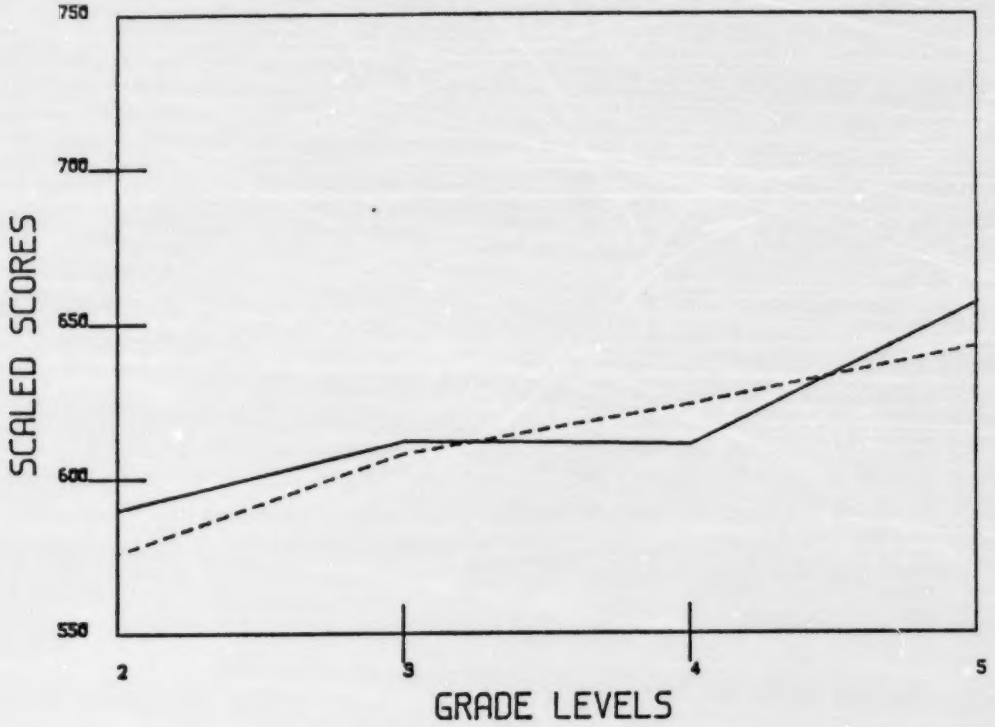
Summary of Results

The speededness of the Reading Comprehension subtest (as evidenced by a district and national omit rate on the final item of 61 and 46 percent, respectively) resulted in some classroom level scores which were substantially different from what would have been expected given class performance on other subtests. Moreover, the completion rate on Reading Comprehension of some classes was noticeably different from what would have been logically predicted from class performance on other subtests. Hence, it appeared that an extraneous factor was operating in the measurement of Reading Comprehension that caused the scores to be inconsistent with other subtests.

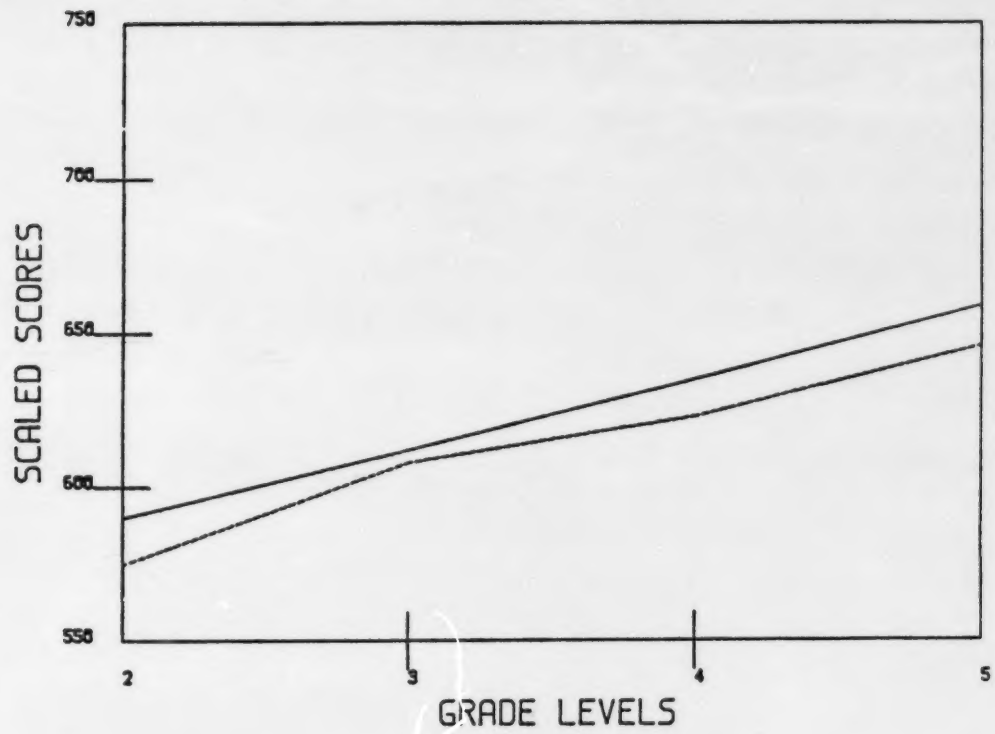
Figure 12

School Profile of Reading Comprehension Scores Across Grades

Reading Comprehension, 60-Item



Reading Comprehension, 40-Item

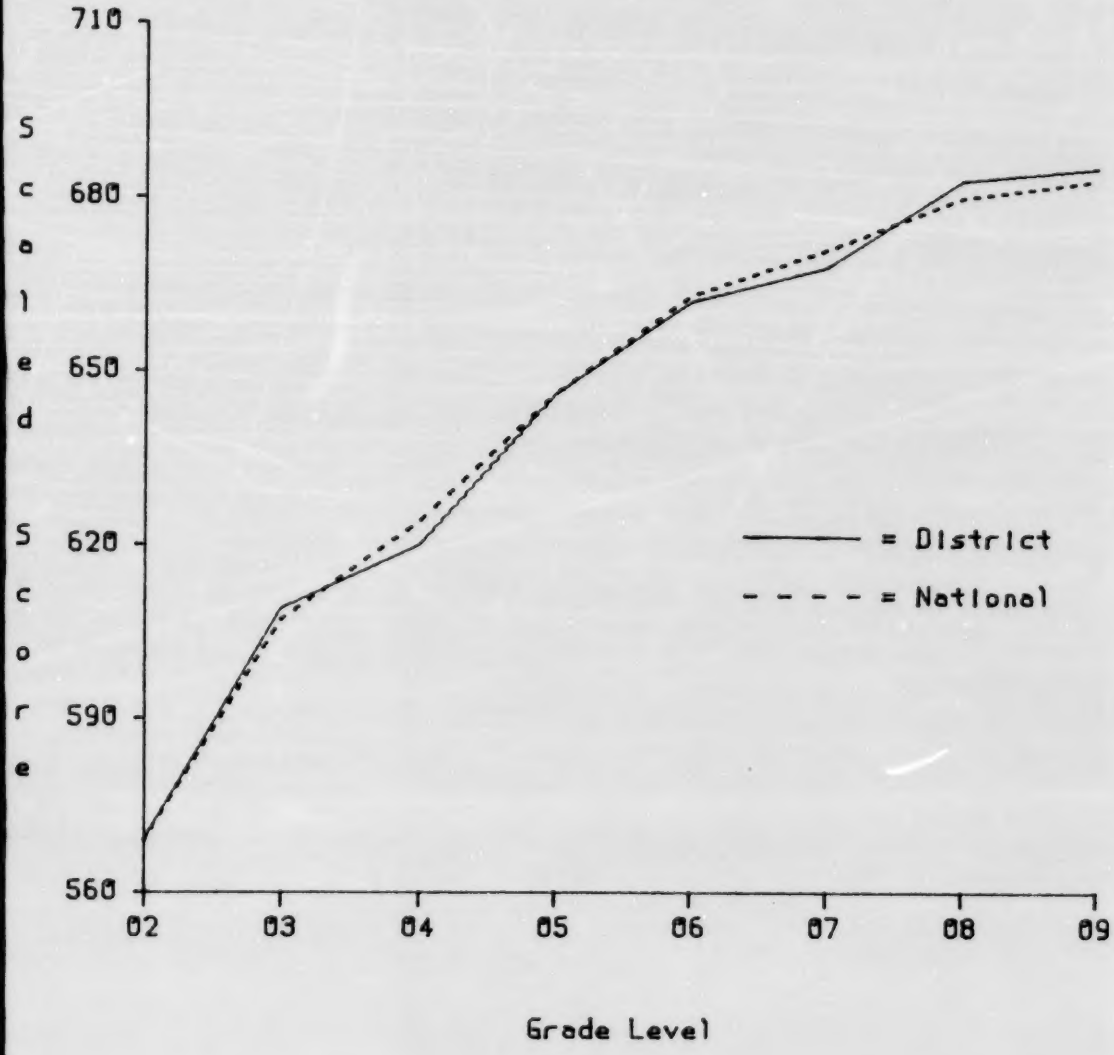


——— = School

. . . . . = National

Figure 13

Reading Comprehension (40-Item) Mean Scaled Scores



As a group, students who completed the Reading Comprehension subtest scored lower on other Stanford subtests than students who completed fewer items. This same result was found to a greater degree when classroom groups were used as the basis of analysis. Since it is reasonable to assume that those students who completed the Reading Comprehension subtest (and those classes with the highest completion rates) would be the highest scoring group on other subtests, it was concluded that some of the students who completed the subtest did so through random guessing, thereby introducing an extraneous factor into the measurement of Reading Comprehension.

Rescoring the Reading Comprehension subtest on the first 40 items with norms developed by the publisher generated scores that alleviated the problems resulting from excessive test speededness. The 40-item score correlated more highly with other subtests for both students and classrooms than did the 60-item score. Additionally, the 40-item scores of initially deviant classrooms were more congruent with the expectations held for those groups. These observations led to the conclusion that the 40-item item score was relatively free from the problems exhibited by the 60-item score.

Replication of the inital investigation at grades 5-7 revealed problems similiar to those found at grade 4. Use of the 40-item scoring at these grades resolved these inconsistencies, as evidenced by school profiles which showed more normal growth curves across grades.

Finally, the district score performance (which was the proximal cause of the investigation) conformed more to expectation when the 40-item score was adopted at grades 4-7.

CONCLUSIONS

The results of the study showed conclusively that an extraneous factor was affecting the Reading Comprehension scores at grades 4-7. It was determined that the highly speeded nature of the Reading Comprehension subtest at these levels of the SAT/7 led to random guessing which resulted in the distortion of that score.

This conclusion brings into question the validity of reading comprehension scores yielded by a highly speeded test. Given that validity can be defined as the interpretive meaningfulness of a test score (Ebel, 1980; Messick, 1980; and the Joint Committee on Standards, 1981), it can be suggested that the 60-item SAT/7 Reading Comprehension subtest administered to students in grades 4-7 in Palm Beach County produced scores of questionable validity. Because the scores were affected to a large extent by random guessing, they could not be interpreted meaningfully.

The findings of this study strongly suggest that any highly speeded measure may be greatly influenced by the intrusion of random guessing. If reading comprehension is to be measured by a speeded instrument, it is essential that the contribution of rate be clearly differentiated from the contribution of power. It is preferred, however, that reading comprehension be measured by a less speeded instrument, and that a separate measure of reading rate be utilized.

BIBLIOGRAPHY

Bloomers, P., & Lindquist, E. F. (1944). Rate of comprehension of reading: Its measurement and its relation to comprehension. The Journal of Educational Psychology, 35(8), 449-473.

Davidson, W. M. N., & Carroll, J. B. (1945). Speed in level components in time limit scores: A factor analysis. Educational & Psychological Measurement, 5, 411-427.

Davis, F. B. (1972). Psychometric research on comprehension in reading. Reading Research Quarterly, 7(4), 628-678.

Ebel, R. L. (1980). Practical problems in educational measurement. Lexington: D. C. Heath.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1981a). Stanford Achievement Test directions for administering (Standardization ed.). USA: The Psychological Corp.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1981b). Stanford Achievement Test, Intermediate I, Form E (Standardization ed.). USA: Harcourt.

Gardner, E. F., Madden, R., Rudman, H. C., Karlsen, B., Merwin, J. C., Collins, R., & Collins, C. S. (1984). Stanford Achievement Test Series: A handbook of instructional strategies. USA: Harcourt.

Helmstadtler, G. C., & Ortmeyer, D. H. (1953). Some techniques for determining the relative magnitude for speed and power components of a test. Educational & Psychological Measurement, 13(2), 280-287.

Joint Committee on Standards for Educational Evaluation. (1981). Standards for evaluation of programs, projects, and materials. New York: McGraw-Hill.

Karlsen, B. (1982a). Diagnosing comprehension: How well can students read what? USA: The Psychological Corp.

Karlsen, B. (1982b). A breakthrough in reading assessment: Stanford. USA: The Psychological Corp.

Kendall, L. M. (1964). The effects of varying time limits on test validity. Educational and Psychological Measurement, 24, 789-800.

Lord, F. M. (1964). The effect of random guessing on test validity. Educational & Psychological Measurement, 24(4), 745-747.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35(11), 1012-1027.

Mollenkopf, W. G. (1960). Time limits and the behavior of test takers. Educational & Psychological Measurement, 20(2), 223-230.

Morrison, E. J. (1960). On test variance and the dimensions of the measurement situation. Educational & Psychological Measurement, 20(2), 231-250.

Myers, C. T. (1960). Introduction. Educational and Psychological Measurement, 20(2), 221-222.

Rindler, S. E. (Winter, 1979). Pitfalls in assessing test speededness. Journal of Educational Measurement, 16(4), 261-270.

Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. Journal of Educational Measurement, 8(4), 275-277.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.) (pp. 356-442). Washington, D. C.: American Council on Education.

Swineford, F. (1956). Technical manual for users of test analyses (Statistical Report 56-42). Princeton, NJ: Educational Testing Service.